Appendix

**A. **Existence Proof of the Optimal Solution for the Optimization Problem:****

**1)Distinction between Decision Variables and Derived Variables**

**Decision variables: Derived variables (determined by decision variables):**

|  |  |
| --- | --- |
|  | (1) |

**Objective function:**

|  |  |
| --- | --- |
|  | (2) |

**2)The feasible region is compact.**

**The decision variables are subject to the following constraints. Quantization-rate and bandwidth-allocation constraints:** *.* **Communication-duration and derived-quantity relationship:**

|  |  |
| --- | --- |
|  | **(3)** |

**Where** 𝑘𝑗 **is a truncated normal random variable, in the optimization it is treated either as a constant or as a deterministic quantity bounded below by** k𝑗≥1 **and is determined simultaneously with** 𝑆𝑗 **Consequently, we obtain** so each 𝑆𝑗 is restricted to a closed and bounded interval as well. Time-shift variable constraints . Link-bandwidth and total-cycle constraints . In summary, all decision variables are confined by a set of linear or bound constraints to a closed and bounded subset 𝑋⊂. By the Heine–Borel theorem, 𝑋 is compact.

3)Continuity of the objective function

Each term / 𝑟𝑗, ∀𝑗, is continuous on the feasible set where the decision variables are strictly positive. The quantity and the operations max (0, ⋅) and min (⋅, ⋅) are point-wise maxima/minima of continuous functions, hence they remain continuous.

|  |  |
| --- | --- |
|  | (4) |

Therefore, the expression is also continuous. Finally, the positive constants 𝛼, 𝛽, 𝛾>0 merely introduce a linear combination and an additional term max (0, 𝑇−∑𝑆𝑗), both of which preserve continuity. Consequently, the objective function 𝑓: 𝑋→𝑅 is continuous everywhere on the compact set 𝑋.

4)Application of the Weierstrass Extreme-Value Theorem

Weierstrass Extreme-Value Theorem:If 𝑋⊆ 𝑅𝑛 is compact and 𝑓: 𝑋→𝑅 is continuous, then 𝑓 attains both its minimum and maximum on 𝑋. From Step 2, 𝑋 is compact; from Step 3, 𝑓 is continuous. Hence 𝑓 must attain its minimum on the feasible region 𝑋 there exists a set of variables , , such that

|  |  |
| --- | --- |
|  | **(5)** |

This completes the proof that an optimal solution to the stated optimization problem necessarily exists under the given constraints.

B. Solution Approach

The objective function is a weighted combination of piece-wise linear terms (involving max, min, and rational expressions). The feasible region is a polytope—i.e., a closed, bounded, convex set defined by linear inequalities. A standard result in convex analysis states that the minimum of a piece-wise linear function over a polytope is attained at one of its **vertices**. Consequently, the global optimummust satisfy , , , in other words, every decision variable is “pinned’’ to an endpoint of its allowable interval.

|  |  |
| --- | --- |
|  | **(2)** |

### *1)* Values of the compression ratio 𝜃𝑗​ and bandwidth 𝑟𝑗

The trade-off between “eliminating overlap penalties” and “minimizing communication time” implies that the optimum is attained at the extremes: . Collectively, only 22𝑁 such combinations need to be enumerated; computing the corresponding 𝑆𝑗 and objective value 𝑓 for each combination and selecting the minimum yields the global optimum.

### *2)Values of the time-shift variable*

Either the communication window of job j is placed flush against the start (or end) of the preceding job j−1 so as to “eliminate” any overlap penalty, or it is pushed to the earliest (0) or latest (T−) position. Hence , wheredenotes the cumulative under a given communication order. This gives four possibilities per job, totaling 4𝑁; Combined with the earlier choices, the overall enumeration is combinations. Physical meaning: Earliest start—communication begins immediately after computation finishes. Window: [,+). Latest start—communication starts exactly at the global latest computation completion time *T*. Window: [T-,T). Right-aligned to the start of the preceding *j*−1 jobs—job *j*’s communication ends then. Window: [,). Left-aligned to the end of the preceding *j*−1 jobs—job *j*’s communication starts then. Window: [,).

3) Closed-form expression of the global optimum

|  |  |
| --- | --- |
|  | **(6)** |

Although the above appears to be “enumeration,” it is exactly the vertex enumeration routine commonly employed for piecewise-linear optimization. The resulting closed-form structure is: and take either their minimum or maximum admissible value; either aligns with a boundary or abuts an adjacent segment.

C. Gradient Quantization Overhead Modeling – Formula Derivation

We model the relative overhead of “gradient quantization + de-quantization” with respect to pure computation time. P: number of model parameters; b: quantization bit-width (bits); : effective hardware throughput of quantization / de-quantization (bits/s); : fixed latency independent of bit-width (ms); , : per-parameter forward and backward compute time (ms/parameter). Quantization + de-quantization latency

|  |  |
| --- | --- |
|  | **(7)** |

Pure computation latency (forward + backward)

|  |  |
| --- | --- |
|  | **(8)** |

Relative overhead

|  |  |
| --- | --- |
|  | **(9)** |

Asymptotic Analysis: Large-model limit (P → ∞). When P dominates, the fixed term becomes negligible:

|  |  |
| --- | --- |
|  | **(10)** |

Small-model regime (finite P) With N small, cannot be ignored:

|  |  |
| --- | --- |
|  | **(11)** |

It can be seen that the overhead of gradient quantization and dequantization relative to forward-and-backward computation time converges, as the number of parameters *P*increases, to a constant ratio determined by the quantization bit‐width b, the hardware throughput , and the per‐parameter computation time.However, when *P* is small, the fixed overhead  remains non‑negligible, causing the actual overhead to exceed this asymptotic bound; thus, special attention must be paid to small‑model scenarios.